

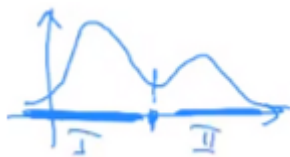
Непараметрические методы оценки плотности

Пусть есть $X_1, \dots, X_n \sim P$ - множество абсолютно непрерывное распределение

Задача - оценить $p: \mathbb{R} \rightarrow \mathbb{R}_+$, $\int p(x)dx = 1$

Зачем?

1. $p \rightarrow E, Var$
2. кластеризация



В книге Wasserman "All of nonparam statistics" есть пример с оцениваем расстоянием до космических объектов. Есть сайт sdss.org. С помощью специальной методики оцениваются расстояния от земли до далеких небесных тел. После этого можно разделить тела на галактики. Это можно сделать построив график плотности. Удалось найти порядка тысячи галактик.

Непараметрическая оценка функции распределения (эмпирическая)

$$F_n(x) = \frac{1}{n} \sum \mathbb{I}\{x_i < x\}$$

она очень хорошая на самом деле.

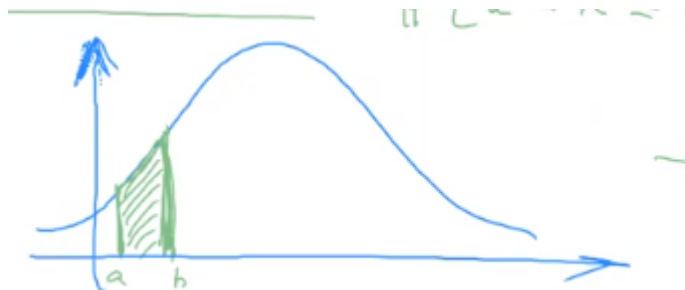
Возникает идея:

$$\hat{p}_n(x) = \frac{d}{dx} \hat{F}_n(x)$$

Другая идея - гистограмма. В математике немного отличается от общепринятого представления

$$P\{a < X \leq b\} = F(b) - F(a) = \int_a^b p(u)du$$

Нужно аппроксимировать интеграл:



Тогда

$$(\dots) = p(x)(b-a), x \in [a, b]$$

(аппроксимировали прямоугольником)

Т.к.

$$P\{x - \frac{h}{2} < x \leq x + \frac{h}{2}\} = p(x)h$$

то

$$p(x) = \frac{P\{x - \frac{h}{2} < x \leq x + \frac{h}{2}\}}{h} \approx \frac{\#\{i : x - \frac{h}{2} < x_i \leq x + \frac{h}{2}\}}{nh}$$

Значит, разбиваем гистограмму на интервалы бины и для каждого считаем статистику

B_1, \dots, B_n — bins

$|B_1| = \dots = |B_n| = h$ — bandwidth

C_1, \dots, C_n — центры

$$B_i = [C_i - \frac{h}{2}, C_i + \frac{h}{2}]$$

$$\hat{p}_n(x) = \frac{\#\{i : x_i \in B_j\}}{nh}, x \in B_j$$

Как оценить качество?

Bias-variance decomposition tradeoff

Первое что приходит в ум — посчитать ошибку

$$\mathbb{E}[\hat{p}_n(x) - p(x)]^2 = MSE(\hat{p}_n(x))$$

Недостаток — зависит от x . Более устойчивая вещь:

$$\int MSE(\hat{p}_n(x)) dx = MISE(\hat{p}_n(x))$$

— mean integer square error

Можно разложить MSE так:

$$MSE(\hat{p}_n(x)) = (Bias(\hat{p}_n(x)))^2 + Var\hat{p}_n(x)$$

$$MISE(\hat{p}_n(x)) = \int (Bias(\hat{p}_n(x)))^2 + \int Var\hat{p}_n(x)$$

В чем компромисс?

Теорема. Если

$$\int (p'(x))^2 dx < \infty$$

то

$$MISE(\hat{p}_n(x)) = (\frac{h^2}{12} \int (p'(x))^2 dx + \frac{1}{nh})(1 + o(1)), n \rightarrow \infty$$

Проанализируем

- первое слагаемое растет как парабола (это bias)
- второе слагаемое убывает как гипербола (это дисперсия)
- есть минимум

Резюме: надо пользоваться MISE

Как выбирается параметр h в R?

Можно просто вычислить ноль производной функции выше:

$$\frac{d}{dh} AMISE(h) = \frac{d}{dh} \left(\frac{h^2}{12} \int (p'(x))^2 dx + \frac{1}{nh} \right)$$
$$h_{opt} = n^{-1/3} \left(\frac{6}{\int p'(x)^2 dx} \right)^{1/3}$$

AMISE - асимптотическая MISE

На самом деле можно было не вычислять производную.

1. $AMISE(h) = f(n)h^{k_1} + f_2(n)h^{k_2}$, $k_1 = 2, k_2 = -1$

Если продифференцировать то получим, что $h_{opt} = C \cdot \left(\frac{f_2(n)}{f_1(n)} \right)^{\frac{1}{k_1+k_2}}$

Но есть фокус: можно просто приравнять порядки:

$$f(n)h^{k_1} = f_2(n)h^{k_2} \Rightarrow \tilde{h} = h_{opt}$$

Это свойство верно и для дальнейших методов

Но мы не можем вычислить эту оценку, т.к. не знаем интеграл. Оценка выше называется оракульной

2. Правило Скотта (Scott's rule):

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-x^2/2\sigma^2\}, N(0, \sigma^2)$$

Отсюда $h_{opt} = n^{-1/3} (24\sqrt{\pi})^{1/3} \sigma \approx 3.5$. Идея:

$$\hat{h}_{opt} = n^{-1/3} 3.5 \hat{\sigma}_n, \hat{\sigma}_n = \frac{1}{n} \sum (x_i - \bar{x})^2$$

3. Правило Фридмана-Дьякони (FD)

$$\hat{h}_{opt} = n^{-1/3} \cdot 2 \cdot \text{IQR}(x_1, \dots, x_n)$$

4. Метод Стерджеса

Пусть у нас Биноминальное распределение: $\xi = \mu_1 + \dots + \mu_n$, $P(\xi = k) = C_n^k p^k (1-p)^{n-k}$.

Нам нужно приблизить эту картинку к нормальному распределению. Если у нас есть n наблюдений, то оптимально взять $m = \log_2 n$.

В языке R все формулы считаются по немного другим формулам

В R используется функция pretty, которая делает число "красивым". Число красивое, если первый знак после запятой это 2, 4, 6, 8, 5 или число вообще целое

Ядерная оценка плотности

density(X)

Kernel density estimate

Kernel $K : \mathbb{R} \rightarrow \mathbb{R}_+$, $\int K(x) dx = 1$, K - четная

- boxcar kernel: $1/2 \mathbb{I}\{|x| < 1\}$
- triangle kernel: $(1 - |x|) \mathbb{I}\{|x| < 1\}$
- Epanechnikov kernel: $3/4(1 - x^2) \mathbb{I}\{|x| < 1\}$
- Gaussian kernel: $\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}$

Kernel density estimate:

$$\hat{p}_n(x) = \frac{1}{nh} \sum K\left(\frac{x - X_i}{h}\right)$$

Можно заметить, что

$$(\dots) \neq 0 \iff \left| \frac{x - X_i}{h} \right| < 1$$

Теорема. Если $\int (p''(x))^2 dx < \infty$, то

$$\text{MISE} = \left(\frac{1}{4} h^4 \int (p''(x))^2 dx \left(\int x^2 K(x) dx \right)^4 + \frac{1}{nh} \int K^2(x) dx \right) (1 + o(1))$$

Приравниваем порядки: $h^4 = 1/nh \Rightarrow h_{opt} \sim n^{-1/5}$.

А с гистограммой мы получили $h_{opt} \sim n^{-1/3}$. Почему другой порядок?

Теорема. Пусть $X_1, \dots, X_n \sim \Phi_m = \{\text{abs. cont.}, \int (p^{(m)}(x))^2\}$

Тогда

$$\sup_{p \in \Phi_m} \text{MISE}(\hat{p}_n) \geq C \cdot n^{-\frac{2m}{2m+1}}$$

Соответственно, для Φ_1 у нас порядок $n^{-2/3}$, для Φ_2 порядок $n^{-4/5}$. То есть гистограммная оценка оптимальна в классе Φ_1 , ядерная - в классе Φ_2 .

Итоговая оценка имеет вид

$$h_{opt} = n^{-1/5} \cdot \left(\frac{\int K^2(x) dx}{\left(\int x^2 K(x) dx \right)^4 \int (p''(x))^2 dx} \right)^{1/5}$$

Можно предположить нормальность распределения и тогда получим оценку

$$\left. \begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \\ K(x) &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \end{aligned} \right\} \Rightarrow h_{opt} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5}$$

В R используется `nrd` для оценки дисперсии:

$$\text{"nrd"} : \sigma \rightarrow \hat{\sigma}_n = \min \left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \frac{IQR(x_1, \dots, x_n)}{1.34} \right)$$

Есть еще аналогичное `"nrd.0"`. Вместо оценки $(4/5)^{1/5}$ берут 0.9 (чисто эмпирически).

Как выбрать ядро?

При фиксированном h мы получаем функционал от ядра, который можно оптимизировать. И для этой задачи даже есть решение

$$h_{opt} = n^{-1/5} \cdot \left(\frac{\int K^2(x) dx}{\left(\int x^2 K(x) dx \right)^4 \int K''(x)^2 dx} \right)^{1/5}$$

$$MISE(\hat{p}_n) = C \cdot n^{-4/5} \cdot \underbrace{\left(\int x^2 K(x) dx \right)^{2/5} \cdot \left(\int K^2(x) dx \right)^{4/5} \cdot \int K''(x)^2 dx}_{\substack{\downarrow \text{argmin} \\ K: \mathbb{R} \rightarrow \mathbb{R}_+, \int K(x) dx = 1 \\ \text{Epanechnikov}}}$$

Ядро Епанечникова - самое лучшее. Но на самом деле выбор ядра не очень важно. Определим эффективность:

$$\text{eff}(K) = \frac{I(K_{ep})}{I(K)} \leq 1$$

Эта штука фигурирует в формуле. На самом деле, почти для всех ядер эффективность близка к 1, просто у Епанечникова она максимальная.

Почему ядро Епанечникова на самом деле не оптимально?

1. Выше была теорема:

$$\text{Теор.} : \int K(x) dx < \infty, \int x^2 K(x) dx < \infty \Rightarrow n^{4/5} \cdot MISE(\hat{p}_n) = \tilde{C}$$

У ядра Епанечникова константа максимальна

2. Посмотрим на ядра в более общем смысле: как функцию из \mathbb{R} в \mathbb{R} .

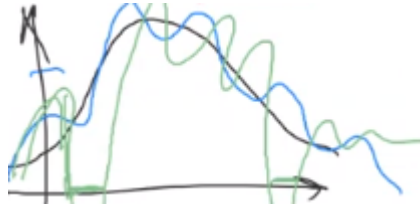
$$K: \mathbb{R} \rightarrow \mathbb{R} \text{ (возм., } < 0), \int K(x) dx = 1$$

$$K \text{ — ядро порядка 2, т.е. } \int x K(x) dx = 0, \int x^2 K(x) dx = 0$$

$$\forall \varepsilon > 0 \exists h = \frac{n^{-1/5} \int K^2(x) dx}{\varepsilon} : \lim_{n \rightarrow \infty} n^{4/5} \text{MISE}(\hat{p}_n) \leq \varepsilon$$

В чем смысл - не понял.

Можно возразить: раньше была хорошая функция, а теперь у нас некрасивая, но более оптимальная. Но есть утверждение:



Утв. теор. означает, что если \hat{p}_n значения то $\hat{p}_n^+ = \max(\hat{p}_n, 0)$

ЕМ алгоритм

- единственный параметрический алгоритм оценки плотности, который хорошо работает

ЕМ - expectation maximization

Рассмотрим смесь двух нормальных распределений:

$$p(x) = (1-\pi) p_{\mu_1, \sigma_1^2}(x) + \pi p_{\mu_2, \sigma_2^2}(x)$$

Параметры и задача:

$$\vec{\theta} = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi)$$

Задача: оценить $\vec{\theta}$ по выборке x_1, \dots, x_n

Первое что приходит в голову - метод максимального правдоподобия

$$\text{ММП: } \sum_{i=1}^n \log p(x_i) := \log L(x_1, \dots, x_n)$$

→ max
 θ

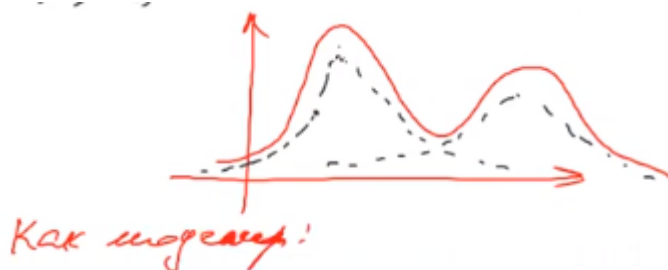
Это сложная оптимизационная задача. Можно использовать метода Бидоля-Равсона. Но это не очень.

Идея

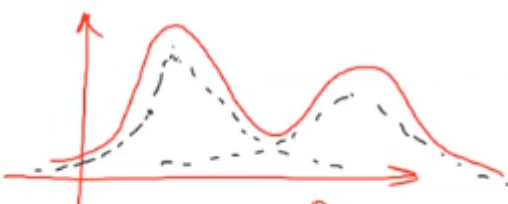
Введем фиктивные (латентные) переменные:

$$Y_i = \begin{cases} 0, & X_i \sim N(\mu_1, \sigma_1^2) \\ 1, & X_i \sim N(\mu_2, \sigma_2^2) \end{cases}$$

У нас распределение является смесью двух других



Отступление: как моделировать величину из смеси распределений



Как моделировать: $\eta = \begin{cases} 0, & 1-\pi \\ 1, & \pi \end{cases}$
 $X \sim N(\mu_1, \sigma_1^2)$, если $\eta = 0$
 $\sim N(\mu_2, \sigma_2^2)$, если $\eta = 1$
 Итог: $X \sim p(x)$

Как доказать? Считаем функцию распределения:

Итог: $X \sim p(x)$

$$P\{X \leq x\} = P\{X \leq x | \eta = 0\} P\{\eta = 0\} + P\{X \leq x | \eta = 1\} P\{\eta = 1\}$$

$\frac{\partial}{\partial x}$: $p(x) = (1-\pi) \varphi(\mu_1, \sigma_1^2)(x) + \pi \varphi(\mu_2, \sigma_2^2)(x)$

Дифференцируем:

$\frac{\partial}{\partial x}$: $p(x) = (1-\pi) \varphi(\mu_1, \sigma_1^2)(x) + \pi \varphi(\mu_2, \sigma_2^2)(x)$

Возвращаемся

Итак, у нас есть Y_i :

$$Y_i = \begin{cases} 0, & X_i \sim N(\mu_1, \sigma_1^2) \\ 1, & X_i \sim N(\mu_2, \sigma_2^2) \end{cases} \quad - \text{заданы } X_i$$

Тогда

$$\begin{aligned} P\{X_i \leq x, Y_i = y\} &= P\{X_i \leq x | Y_i = y\} P\{Y_i = y\} = \\ &= \begin{cases} \Phi(\mu_1, \sigma_1^2)(x) \cdot (1-\pi) & , y=0 \\ \Phi(\mu_2, \sigma_2^2)(x) \cdot \pi & , y=1 \end{cases} = \\ &= (1-y)(1-\pi) \Phi(\mu_1, \sigma_1^2)(x) + y\pi \Phi(\mu_2, \sigma_2^2)(x) \quad \left(\frac{\partial}{\partial x}\right) \end{aligned}$$

Дифференцируем и приводим к красивому виду:

$$\begin{aligned} p_{X,Y}(x,y) &= (1-y)(1-\pi) \phi(\mu_1, \sigma_1^2)(x) + y\pi \phi(\mu_2, \sigma_2^2)(x) = \\ &= ((1-\pi) \phi(\mu_1, \sigma_1^2)(x))^{1-y} \cdot (\pi \phi(\mu_2, \sigma_2^2)(x))^y \end{aligned}$$

Записываем лог-правдоподобие:

$$\begin{aligned} p(x) &= (1-\pi) \phi(\mu_1, \sigma_1^2)(x) + \pi \phi(\mu_2, \sigma_2^2)(x) \\ \text{ММП: } \sum_{i=1}^n \log p(x_i) &:= \log L(x_1, \dots, x_n) \end{aligned}$$

$$\begin{aligned} \log L_\theta(x_1, \dots, x_n, y_1, \dots, y_n) &= \sum_{i=1}^n \log p_{X,Y}(x_i, y_i) = \\ &= \sum_{i=1}^n (1-y_i) \cdot \log(1-\pi) + \sum_{i=1}^n (1-y_i) \log \phi(\mu_1, \sigma_1^2)(x_i) \\ &\quad + \sum_{i=1}^n y_i \log \pi + \sum_{i=1}^n y_i \log \phi(\mu_2, \sigma_2^2)(x_i) \end{aligned}$$

Но как отсюда вытащить оценку? Запишем $\log L_\theta(x_1, \dots, x_n, y_1, \dots, y_n)$ и вместо y_i подставим Y_i - случайные величины. А потом возьмем матожидание и максимизируем по θ

$$\mathbb{E}_\theta [\log L_\theta(x_1, \dots, x_n, Y_1, \dots, Y_n) | X_1 = x_1, \dots, X_n = x_n] \rightarrow \max_{\theta}$$

Как выглядит алгоритм

1. Выбираем $\theta^{(0)}$

2. E-step

$$\begin{aligned} \tilde{E}\text{-step} \quad E_{\theta} [\log L_{\theta}(x_1, \dots, x_n, y_1, \dots, y_n) | x_1 = x_1, \dots, x_n = x_n] &= \\ \sum_{i=1}^n (1 - e_i) \log(1 - e_i) + \sum_{i=1}^n (1 - e_i) \log p_{\mu_i, \sigma_i^2}(y_i) + \dots \end{aligned}$$

Тут

$$e_i = E_{\theta}[y_i | x_i = x_i] = P\{y_i = 1 | x_i = x_i\}$$

Вспомним формулу Байеса:

$$P\{A|B\} = \frac{P\{B|A\}P\{A\}}{\sum P\{B|A_k\}P\{A_k\}}$$

Тогда

$$e_i = E_{\theta}[y_i | x_i = x_i] = P\{y_i = 1 | x_i = x_i\} = \frac{P(\mu_i, \sigma_i^2)(x_i) \cdot \pi}{P(\mu_i, \sigma_i^2)(x_i) \cdot \pi + P(\mu_i, \sigma_i^2)(x_i) \cdot (1 - \pi)}$$

3. M-step:

4. Вернуться к шагу 1

Зачем нужны были латентные переменные? Они постоянно убираются

Почему это работает?

Изначальная идея - построить оценку лог правдоподобия

$$\log L_{\theta}(x_1, \dots, x_n) \rightarrow \arg \max_{\theta}$$

Преобразуем через условные вероятности:

$$\sum_{i=1}^n \log p_{\theta}(x_i) = \sum_{i=1}^n \log \frac{p_{\theta}(x_i, y_i)}{p_{\theta}(y_i | x_i)}$$

и распишем логорифм:

$$= \sum_{i=1}^n \log p_{\theta}(x_i, y_i) - \sum_{i=1}^n \log p_{\theta}(y_i | x_i)$$

Т.к. y_i случайные, можно взять любые. Возьмем случайные y_i , то есть Y_i и используем матожидание:

$$\log \zeta(x_1, \dots, x_n) = \mathbb{E}_{\theta} [\log \zeta(x_1, \dots, x_n, y_1, \dots, y_n) | x_1 = x_1, \dots, x_n = x_n] - \mathbb{E}_{\theta} [\sum \log p_{\theta}(y_i | x_i) | x_1 = x_1, \dots, x_n = x_n]$$

Посмотрим, как у нас меняется логправдоподобие после итерации:

$$\log \zeta_{\theta^{n+1}} - \log \zeta_{\theta^n} = \underbrace{\mathbb{E}_{\theta^n} [\log \zeta_{\theta^{n+1}} \dots] - \mathbb{E}_{\theta^n} [\log \zeta_{\theta^n} \dots]}_{> 0} + \underbrace{\mathbb{E}_{\theta^n} [\sum \log \frac{p_{\theta^{n+1}}(y_i | x_i)}{p_{\theta^n}(y_i | x_i)} | x_1 = x_1, \dots, x_n = x_n]}_{> 0}$$

Второе слагаемое это KLD:

Kullback-Leibler divergence > 0

$$K(f, g) = \mathbb{E}_f \left[\log \frac{f(x)}{g(x)} \right] = \int f(x) \log \frac{f(x)}{g(x)} dx$$

Получается, мы с каждым шагом работы EM алгоритма увеличиваем логорифм правдоподобия, а значит и приближаемся к ней. Возможно, мы попадем в локальный минимум, но в любом случае движемся мы в правильном направлении